INFORMATION AND CONTROL 4, 300–304 (1961)

Post Scriptum to "Final Note"

BENOIT MANDELBROT

I.B.M. Research Center, Yorktown Heights, New York

My criticism has not changed since I first had the privilege of commenting upon a draft of Simon (1955).

I. REMARKS ON THE FIRST ASSUMPTION[1]

Professor Simon's reference to his Assumption I implicitly concedes all my points, because, as should have been obvious all along, every important prediction based upon Assumption I′ also applies in the case of Assumption I. The *mathematical* equivalence between the two extends to all the items (words, cities, incomes, . . .) for which the *actual* value of $f(i, k)$ is equal to 1 or 0. It is quite easy to see that this holds roughly for all $i > i'$, where the *expected* value $f(i', k)$ (as given by the theory) is equal to 1. That is, the two assumptions fully agree for a number of genera, words, cities, incomes, . . . that increases without bound with k. For example, taking reasonable values of k and ρ, we find that this would apply to the 10^3 most frequent words. The mathematical prediction that $f(i, k) = 0$ or 1 for the most frequent items is of course empirically borne out by the fact that the sizes of the 50-odd largest genera are all different (see Yule, 1924) as are the sizes of the 300 USA cities given in my copy of the Almanac.

Hence, Simon's theory indeed fails for cities and incomes because of the

[1] The answers to Simon's objections are given in the following sections.

I.1: see Section I. **I.2:** see Section II. **II.1:** see Section IV, first half. **II.2:** see Section IV, second half **II.3:** see Section IV, first half. **II.4:** see Sections I and III. **III.1:** see Section III. **III.2:** we are concerned with carrying out a model, before looking at the data; for $1 < \rho < 1.1$ see Section II. **IV:** it served to get Simon to define "slowness." **V.1:** see Section II. **V.2:** see Section IV, second half. **V.3:** same conclusion if ρ varies from 0.95 to 0.05. **V.4:** see Section III; this is a crucial point. **V.5:** in our own words, "a geometric distribution is not concentrated around any particular value," what we approximate by a gaussian is the negative binomial for large k and $i°$. **VI.1:** we still do not see any mathematical errors—on our side of the fence, that is—. **VI.2:** see Section V; we do not understand the second sentence. **VI.3:** see end of Section II. **VI.4:** see end of Section V. **VI.5:** the bias is infinitesimal and the data on $n(k)$ are nonexistent.

inadequacy of its diffusion term. It also fails for word frequencies because with constant $n'(k)$ and ρ larger than 1.1 (say), the probabilities i/k of *all* the 10^3 most frequent words (and not only of the verb "art") rapidly become ridiculously low. (The other linguistic cases will be treated below.)

To eliminate these difficulties, it is necessary to completely modify the structure of Simon's model, to obtain something like the diffusion process implicit in Champernowne's random walk of log i.

II. REMARKS ON VERY SLOWLY DECREASING $n'(k)$

We are glad that Simon has agreed to narrow his claims to the case of "*very* slowly decreasing" $n'(k)$, defined as being such that $\sum^{\infty} n'(k)/k$ diverges. (He says that he limits himself to "$n'(k)$ like those encountered in the data" and, elsewhere, that he "is persuaded that the convergence of $\sum_1^{\infty} n'(h)/h$ is the exception." Note that this reference to "data" is of course without basis, since one can only observe the function $d(k)$ defined in Section VI of our "Final Note," while Simon's models are based upon the function $n(k)$ relative to first occurrences of words in a child's speech.)

Practically, the case of very slow decrease can be combined with the case where $n'(k)$ is constant, but ρ is less than 1.1 (say), so that the decrease of i/k for all words is acceptably slow. In both cases, log $n(k)$, considered as a function of log k, will asymptotically parallel the main diagonal, the elasticity of $n(k)$ being asymptotically one. We are surprised to see Simon invoke such asymptotics, as we thought that he wanted approximations valid for all i. But it is quite true that one can prove that if the asymptotic elasticity of $n(k)$ is one, then log (rank), considered as a function of log i, will be *asymptotically* parallel to the second diagonal. This is a particular example of analytic circularity, in which the formal identity of "input" and "output" is only asymptotic.

In accepting this definition of very slow decrease Simon has of course excluded all $\rho < 1$, for which all hands agree that models are required.

III. REMARKS ON THE DERIVATION OF $f(i,k)$

We have shown how to derive $f(i, k)$ from $n'(k)$ by an integral transform (2.6), to which equation (7) of Simon's reply is essentially identical. This has been accepted; but our division of the transformation into two steps was not understood. Let us repeat: when one considers increments Δk small compared to k^0, one may safely neglect the very few words in that sample that, in the whole history of the process, have occurred less than i^0 times; say, less than 10 times. Hence, practically all the words will have sometime passed through the value i^0 of i. We never claimed that the i cluster around i^0. The Chapman-Kolmogoroff equation is irrelevant to our problem.

However, $f(i^0, k)$ is a i^0 times integrated and *much smoothed-out* form of $n(k)$. This shows that the smoothing from $n(k)$ to $f(i, k)$ (which is, even so, inadequate to support Simon's original claims) results from the application of Assumption I or I′ to *very small* values of i, before a word had a chance of becoming part of any permanent verbal system. I or I′ seem reasonable because of our feeling for the proportionality of Δi to Δk; but this does not make them any less uncheckable, crude, and arbitrary for small i. One really should not hope to be able to give stochastic models of such phenomena in this way, and one really cannot trust any conclusions drawn from such models. Even with I as applied to very small i, one cannot get the right form for $f(i, k)$; without I, one gets any locally smooth decreasing function.

Actually, our point about small i was implicitly accepted by Simon, since it is the same as his comment, on p. 438 of (1955), that his model "could only be expected to hold to some minimum city size, say 5000 or 10000." (Similarly, his model could not apply to small values of income —to which the law of Pareto is anyway inapplicable.) If i^0 is taken to be 10000, the root mean square of the relative fluctuation of $i/H(k)$ is $100(10000)^{-1/2} = 1\%$ and the size distribution of cities, first generated by the mechanism valid for $i < 10000$, will be carried on forever. That is, the model turns out again to be completely undeterminate.

The above argument does not apply to the Willis relationship. But, in that case, the representation of $f(i, k)$ by $Ci^{-(\rho+1)}$ is so good that one may without fear invert the Laplace transform, to obtain the requirement $n(k) \sim Ck^{\rho}$.

IV. REMARKS CONCERNING SIMON'S MATHEMATICS

We fail to see the point of Simon's variant for our derivation of the integral transform that gives $f(i, k)$. His *unconscious* approximation of k by a continuous variable leads him directly to the questionable form of $f(i, k)$ in which i is unbounded. He adds nothing by using a different notation, $\tau(k) = \log\,[H/(k)H(k^0)]$, in which k^0 does not appear explicitly; besides, $\tau(k) \sim \log\,(k/k^0)$, as he claims, only in the case when

$\sum^{\infty} n'(k)/k$ converges, which is supposed to be excluded by "the data".

As to the much repeated claims for the expression

$$\sigma(k) = kn'(k)/n(k)[1 - n'(k)]$$

they are too absurd, even when one disregards the worries about the infinitesimal "bias" which would be brought by the deletion of the term $1 - n'(k)$. The transform that yields $f(i, k)$ is obviously taken for fixed k, but it involves $n'(h)$ for *all* $h \leq k$. Moreover, one is interested in the rank-frequency function relative to an increment Δk small with respect to k (say, $\Delta k \sim 10^6$, while $k \sim 10^9$); this distribution and the corresponding value of ρ (if it exists) are for practical purposes entirely independent of the current $n'(k)$ or $\sigma(k)$, and are fully determined by *long past* values of $n'(h)$. There is no sense in speaking of a variation of ρ with Δk—and no evidence for it. In other words, anyway, $\sigma(k)$ is a *local* property of $n(k)$ and, unless it is constant so that the elasticity of $n(k)$ is constant, $\sigma(k)$ cannot possibly describe any global properties of $f(i, k)$, such as the slope of the straight line which best approximates log $f(i, k)$ as a function of log i (this best approximation is usually likely to be pretty bad).

V. REMARKS ON IMPROPER DISTRIBUTIONS AND ON THE ESTIMATION OF A CERTAIN ρ

We are relieved to learn that distributions are no longer "improper," "because" their moments are infinite, but because their range of values is infinite. Their prototype is therefore the gaussian and one needs not be too concerned. Incidentally, my mathematical error, pointed out on p. 84 of Simon (1960) and recalled on p. 222 of his (1961), refers to the use of improper distribution functions.

To check the actual extent of the breakdown of $f(i, k)$, as a representation of the data for large i, it is best to examine the rank-frequency law, which should take the form $r(i, k) = -V + Ci^{-\rho}$, where the term V puts together the fluctuations and the imperfections of the representation. Of course, this works for word frequencies, with a V that is small and varies little (in our theory, V has a different origin).

But what about species-genera data, such as those of the insert of 9.7 in Zipf (1949)? The curve log $f(i, k)$, which Zipf gives, is remarkably straight with a slope between 1.5 and 1.6. As a result, the value of ρ is contained between the limits 0.5 and 0.6, as we have said. But, here, for some reason, Simon prefers to evaluate ρ from the rank-frequency curve, with $V = 0$, hereby *implying* for once *that* $f(i, k)$ *applies to unbounded* i. He finds that log $r(i, k)$ is not straight and that its "local" slope goes from 2 to 1, but never to $\frac{1}{2}$. We were unfortunately unable to check whether this is due to an exceptionally fast breakdown of the Pareto law for large i or *rather* a consequence of Zipf's habitual omission, from his number-frequency graphs, of most or all of the values of i that occur only once. For example, see Zipf's figure 2.3 (truncated to words appearing less than 50 times) or his figure 3.3 (truncated precisely to words for which $f(i, k)$ is greater than 1) or his figure 3.4 (truncated in the same manner; this graph also gives the rank-frequency curve and it is obvious that this curve could not have been deduced from the portion of $f(i, k)$ reported on this graph). Hence, none of the empirical counter-examples of Simon's can be maintained.

We hope that the question is now settled. We had no need for any numerical simulation method.

References

Simon, H. A., (1960). Some further notes on a class of skew distribution functions. *Information and Control* **3,** 80 (1960).

Zipf, G. K. (1949). "Human Behavior and the Principle of Least Effort." Addison Wesley, Reading, Massachusetts.